
Proceedings Paper:
Picardi, Chiara and Habli, Ibrahim orcid.org/0000-0003-2736-8238 (2019) Perspectives on Assurance Case Development for Retinal Disease Diagnosis Using Deep Learning. In: AIME 2019: Artificial Intelligence in Medicine. Lecture Notes in Artificial Intelligence . Springer , pp. 365-370.

https://doi.org/10.1007/978-3-030-21642-9

Perspectives on Assurance Case Development for Retinal Disease Diagnosis using Deep Learning

Chiara Picardi and Ibrahim Habli

University of York, York YO10 5DD, UK
{chiara.picardi,ibrahim.habli}@york.ac.uk

Abstract. We report our experience with developing an assurance case for a deep learning system used for retinal disease diagnosis and referral. We investigate how an assurance case could clarify the scope and structure of the primary argument and identify sources of uncertainty. We also explore the need for an assurance argument pattern that could provide developers with a reusable template for communicating and structuring the different claims and evidence and clarifying the clinical context rather than merely focusing on meeting or exceeding performance measures.

Keywords: Assurance case · Machine learning · Retinal disease · Safety

1 Introduction

Justifying the use of machine learning in critical healthcare applications is currently a significant technological and societal challenge [5]. The developers and clinical users of the technology have to assure, prior to deployment, different critical properties such as safety, performance, usability and cost-effectiveness [6]. This challenge can be refined further into 2 parts. Firstly, there is no consensus on the assurance criteria or specific properties that machine learning systems have to exhibit for them to be accepted by the public or by the clinical and regulatory authorities, i.e. what is good enough? Secondly, there is very little guidance, e.g. standards, on accepted means for achieving such properties [6].

In this paper, we investigate the extent to which an explicit assurance case could inform a decision concerning the use of machine learning in clinical diagnosis. An assurance case is *"a reasoned and compelling argument, supported by a body of evidence, that a system, service or organisation will operate as intended for a defined application in a defined environment"* [1]. An assurance case is considered as a generalisation of a safety case, i.e. where safety claims are the focus of the assurance.

We build on the results of De Fauw et al [2] on the use of a deep learning system for diagnosis and referral in retinal disease. This system comprises 2 different neural networks. The first network, called Segmentation Network, takes as input three-dimensional Optical Coherence Tomography (OCT) scans and creates a detailed device-independent tissue-segmentation map. The second network examines the segmentation map and outputs one of the four referral suggestions in addition to the presence or absence of multiple concomitant retinal pathologies.

Through an assurance case, our objectives are to (1) clarify structure of the primary argument and the clinical context and (2) identify sources of uncertainty. The contribution of the paper is that it provides a self-contained assurance case for a deep learning system, thereby highlighting assurance issues that have to be considered explicitly beyond merely exceeding a specific performance measure.

2 Assurance Case

The assurance case is represented using the Goal Structuring Notation (GSN) [1]. GSN is a generic argument structuring language that is widely used in the safety-critical domain. The reader is advised to consult the publicly available GSN standard [1] for a more detailed description of the notation. Due to the space limitation, we focus the discussion on 2 assurance argument fragments:

1. Segmentation network assurance argument (Figure 1, Section 2.1)
2. Classification network assurance argument (Figure 2, Section 2.2)

These arguments capture the essence of the justification based on performance against clinical experts. The clinical context in the assurance case is the ophthalmology referral pathway at Moorfields Eye Hospital, from which the training, validation and test data is provided. At this stage, the scope of the claims is limited to this clinical setting with no evidence for generalisation (despite the wide and diverse population served). It is important to note that the assurance case focuses exclusively on the chain of reasoning and evidence based on the data in the original study [2]. The extent to which this assurance case could be improved, or its scope extended, is discussed in Section 3.

2.1 Segmentation Network Assurance Argument

Figure 1 shows the assurance argument fragment concerning the performance and transparency of the segmentation network. The argument makes a distinction between the scans that include ambiguous and unambiguous regions. The context is important here, referencing the data used for training, testing and validation. It also clarifies the profile of the clinical experts involved in the segmentation experiment. Evidence of sufficient performance is provided based on two different scanning devices (99.21% and 99.93%). The argument clarifies further that for unambiguous regions, the network produces tissue-segmentation maps that are comparable to manual segmentation. For scans with ambiguous regions, the network provides different (but plausible) interpretations of the low quality regions, i.e. similar to how the different human experts might produce different interpretations. The evidence is represented by supplementary videos that show the multiple hypotheses of the segmentation maps produced by the network. An important aspect of creating a separate network for segmentation is greater transparency. By being able to inspect the tissue-segmentation map (and not just referral decisions), clinicians have clearer means for understanding the basis for the final clinical decision. What is less clear, however, is the

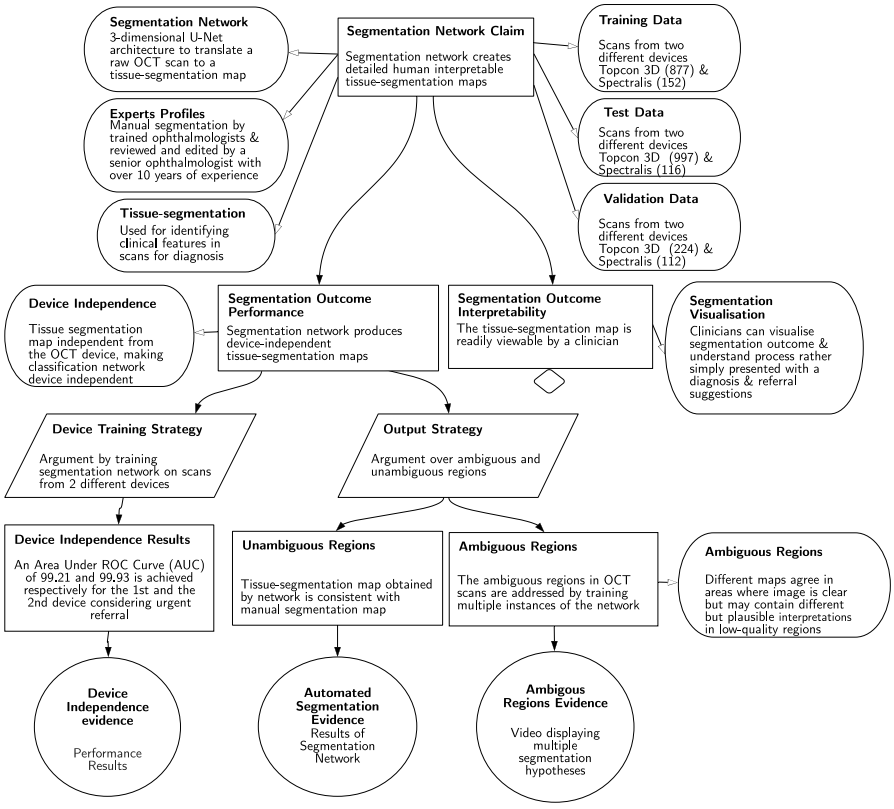

Fig. 1. Segmentation Network Assurance Argument

effectiveness of this visualisation, i.e. degree of acceptance by clinical experts. As such, this is labelled as *'to be developed'* (small diamond below the claim).

2.2 Classification Network Assurance Argument

The argument in Figure 2 states the primary claim that the system achieves or in some cases exceeds human expert performance in retinal disease diagnosis and referral. Experts comprise 4 retina specialists with respective 21, 21, 13 and 12 years of experience and 4 optometrists with respective 15, 9, 6 and 3 years of experience. Two sessions were organised. In the first session experts were required to give the referral suggestions using the OCT scans only. In the second session they were also able to use fundus images and clinical notes. Similar to the segmentation network assurance argument, this argument communicates clearly the training, test and validation data as well as the benchmark against which performance is assessed (i.e. gold standard and expert profiles).

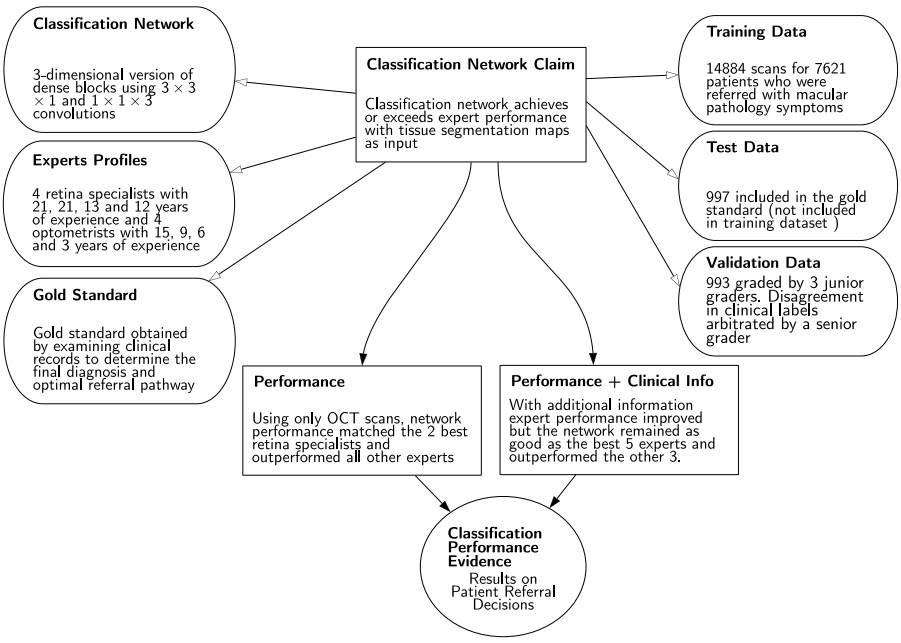

Fig. 2. Classification Network Assurance Argument

3 Discussion

We reflect on the insights gained and lessons learned from different perspectives.

Performance-based Arguments. Evidence in machine learning studies tends to focus on meeting or exceeding certain performance criteria. The assurance argument above is consistent with this approach. Importantly, it ensures that the different training, test and validation datasets are explicitly referenced in addition to the performance results. It clarifies, particularly to non-technical reviewers and decision makers, the importance of appraising the quality of these datasets and the extent to which the data used is relevant to the context in which the performance claims are made. The argument also prompts the reviewers to question the performance criteria used.

Assurance Case Pattern. By looking at the argument fragments for the Segmentation and Classification Networks, a *pattern* of reasoning seems to emerge (Figure 3). Such a pattern could prompt the developers and assessors of machine learning to more explicitly consider the relevance and appropriateness of the contextual and evidential data, i.e. ensuring sufficient confidence in the quality and relevance of the data and models, by scrutinising the *links* in the argument in Figure 3, rather than merely exceeding a specific performance measure.

Assumptions and Transparency. An assurance case can help ensure that the assumptions made are explicitly listed. For example, the reviewers of the case

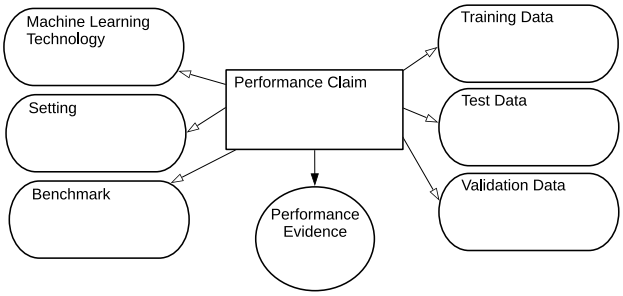

Fig. 3. Preliminary Machine Learning Assurance Argument Pattern

can question the profiles and representativeness of the clinical experts involved in the experiments and the extent to which further clarification might be necessary. Transparency in how the machine makes clinical decisions is also important. Here, the assurance case clarifies that transparency is limited to the output of the segmentation and not the classification network, i.e. prompting the reviewer to question the need for transparency in the final diagnosis and referral decision.

Safety and Regulations. Although our assurance case does not directly address patient safety [3], there remain fundamental questions as what is deemed as good enough for assuring the safety of machine learning. For example, are arguments based on exceeding human equivalence or appealing to risk-benefit evidence acceptable? How do we address non-quantifiable factors such as those related to human or organisational factors? Another issue is the readiness of the regulators to review, challenge and approve machine learning evidence. Kelly in [4] talks about the *Imbalance of Skills* between the developers and the independent assessors of novel technologies as a major hurdle for effective assurance case practices. The readiness of regulators to appraise machine learning algorithms, evaluation evidence and deployment constraints is an ongoing concern.

References

1. Assurance Case Working Group [ACWG]. Goal structing notation community standard version 2. https://scsc.uk/r141B:1?t=1, 2018. (Accessed on 11/13/2018).
2. Jeffrey De Fauw et al. Clinically applicable deep learning for diagnosis and referral in retinal disease. *Nature medicine*, 24(9):1342, 2018.
3. Ibrahim Habli, Sean White, Mark Sujan, Stuart Harrison, and Marta Ugarte. What is the safety case for health it? a study of assurance practices in england. *Safety Science*, 110:324–335, 2018.
4. Tim Kelly. Are safety cases working. *Safety Critical Systems Club Newsletter*, 17(2):31–33, 2008.
5. Thomas M Maddox, John S Rumsfeld, and Philip RO Payne. Questions for artificial intelligence in health care. *Jama*, 2018.
6. Edward H Shortliffe and Martin J Sepúlveda. Clinical decision support in the era of artificial intelligence. *Jama*, 2018.